Environ Monit Assess (2015) 187: 694
DOI 10.1007/s10661-015-4901-z

Wavelets-based clustering of air quality monitoring sites

Sónia Gouveia · Manuel G. Scotto · Alexandra Monteiro · Andres M. Alonso

Received: 6 March 2015 / Accepted: 29 September 2015 / Published online: 19 October 2015
© Springer International Publishing Switzerland 2015

Abstract This paper aims at providing a variance/covariance profile of a set of 36 monitoring stations measuring ozone (O_3) and nitrogen dioxide (NO_2) hourly concentrations, collected over the period 2005–2013, in Portugal mainland. The resulting individual profiles are embedded in a wavelet decomposition-based clustering algorithm in order to identify groups of stations exhibiting similar profiles. The results of the cluster analysis identify three groups of stations, namely urban, suburban/urban/rural, and a third group containing all but one rural stations. The results clearly indicate a geographical pattern among urban stations, distinguishing those located in Lisbon area from those located in Oporto/North. Furthermore, for urban stations, intra-diurnal and daily time scales exhibit the highest variance. This is due to the more relevant chemical activity occurring in high NO_2 emissions areas which are responsible for high variability on daily profiles. These chemical processes also explain the reason for NO_2 and O_3 being highly negatively cross-correlated in suburban and urban sites as compared with rural stations. Finally, the clustering analysis also identifies sites which need revision concerning classification according to environment/influence type.

Capsule abstract: Wavelet NO_2 and O_3 profiles obtained for Portuguese monitoring stations point out the need to further review their classification according to environment/influence type.

S. Gouveia (✉)
Instituto de Engenharia Electrónica e Informática de Aveiro (IEETA) and Centro de I&D em Matemática e Aplicações (CIDMA), Universidade de Aveiro,
Campo Universitário de Santiago,
3810-193 Aveiro, Portugal
e-mail: sonia.gouveia@ua.pt

M. G. Scotto
CEMAT, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

A. Monteiro
Centre for Environmental Marine Studies (CESAM) and Department of Environment and Planning, Universidade de Aveiro, Aveiro, Portugal

A. M. Alonso
Department of Statistics and INEACU, Universidad Carlos III de Madrid, Madrid, Spain

Keywords Air quality monitoring stations · Ozone · Nitrous oxide · Wavelets · Classification · Clustering

Introduction

Air quality monitoring in specific locations is the main tool for governmental management and evaluation of local air quality status. The classification of monitoring stations follows technical regulation (type of environment and of influence) and highlights the similarities among sites with respect to site characteristics, pollutant concentration levels, and/or temporal profiles. This procedure also allows to disclosure regional

patterns by the aggregation of stations with similar profiles. However, stations should not be assumed to be classified correctly once updates and changes in environment conditions are likely to occur. Additionally, the classification of stations can become very problematic since legislation requires the measurement of multiple air pollutants (Council Decision 97/101/EC on Exchange of Information) which complicates the interpretation, and especially, the global integration of all information that defines air quality (e.g., Monjardino et al. 2009).

Methodologies for station categorization and for classification accuracy assessment are typically based on the statistical analyses of concentrations records following by a clustering procedure aiming at grouping stations with similar profiles (e.g., Joly and Peuch 2012; Kracht et al. 2013, 2014; Sharma and Kulshrestha 2014). Within this framework, the variability of the concentrations records has received much attention (e.g., Levy et al. 2014; Li et al. 2013; O'Leary and Lemke 2014).

In studies of regional pollutant concentrations variability, air quality records can be analyzed individually (e.g., Figueiredo et al. 2013; Reich et al. 2013; Carvalho et al. 2010; Adame et al. 2010; Rojas and Venegas 2013) or simultaneously (e.g., Finazzi et al. 2013; Im et al. 2013; Shaddick and Wakefield 2002; Clapp and Jenkin 2001). An alternative multivariate approach is to consider, simultaneously, the whole data set of pollutant concentrations measurements for each station and characterize regional variability by means of techniques such as empirical orthogonal functions (e.g., Alkuwari et al. 2013; Fiore et al. 2003), Canonical Correlation Analysis (e.g., De Iaco 2011; Statheropoulos et al. 1998), and cross-correlation analysis (e.g., Monteiro et al. 2012a). To this extent, cluster analysis provides a powerful tool for characterizing regional variability in terms of locations exhibiting similar patterns. However, although clustering techniques have been widely popular for the analysis of non-time series environmental data, its extension to time series data is hindered by the serial dependence and high-dimensionality of the observations. Despite this fact, clustering of time series is a rapidly developing subject and it has been a topic of active research over recent years, mainly due to its wide applicability to the analysis of environmental processes. In particular, cluster analysis of time series pollutant concentrations has also received much attention in the literature. A comprehensive revision of ozone-based clustering approaches can be found in Ignaccolo et al. (2008) and references therein. More recently, D'Urso et al. (2014) applied the wavelet-based clustering approach proposed by D'Urso and Maharaj (2012) in the analysis of pollutant concentrations (CO, NO, and NO_2) in Rome, Italy. Monteiro et al. 2012b applied a clustering procedure based on quantile regression to explore the changes in hourly O_3 data collected over the Iberian Peninsula from 2000 up to 2009. Alonso et al. (2006) introduced a forecast-density-based time series classification method and analyzed historical data of CO_2 emissions in industrialized countries. Extensions of the clustering procedure proposed by Alonso et al. (2006) have been proposed by Vilar et al. (2010). The authors considered non-parametric approximations to the true autoregressive functions without making any assumptions on parametric models for the true autoregressive structure of time series; see also Liu et al. (2014) for further extensions. Also, Shi et al. (2014) employ the k-means clustering algorithm to classify the daily variation cycles of SO_2 and NO_2 recorded at the north of Xiamen (China) considering the temperature, relative humidity, wind speed, and direction as covariates.

It is well-known that air quality time series exhibit several periodicities with different contributions to the total variance of the series (Sebald et al. 2000; Tchepel et al. 2010). For example, spectral analysis points out that the daily period (24 h) has the largest importance in the hourly O_3 time series (Hogrefe et al. 2000). Additional frequency bands of interest are intra-day (periods less than 12 h expressing local-level processes, high frequency), synoptic (periods of 2–21 days associated with changing weather patterns, intermediate frequency), and longer-term (i.e., baseline that contains longer periods including the yearly periodicity reflecting changes over the entire year, low frequency). Thus, the relative contribution of these periodicities to the total variance/covariance of a set of multivariate time series can be used to build a profile to characterize the monitoring location in terms of variance/covariance decomposition.

The goal of this work is to build a variance/covariance profile for Portuguese monitoring stations across time scales with direct connection with the main periodicities observed in NO_2 and O_3 time series (2005–2013). The individual profiles are then

used to identify groups of similar stations and also to investigate the classification (environment and influence) of monitoring sites. The stations profile is based on a discrete wavelet variance/covariance analysis since NO_2 and O_3 time series are known to be of non-stationary nature with important changes in time of major periodicities. This ruled out the possibility of considering standard power spectrum techniques such as the well-known fast Fourier transform (FFT) in the analysis of such time series. Wavelet variance provides the scale-by-scale variance of the univariate time series (NO_2 and O_3) while the wavelet covariance provides the scale-by-scale joint covariance (or association) between each pair of wavelet scales. Therefore, distance measures based on such features will provide detailed variance/covariance information about the time series at different frequency intervals. Note that such information cannot be obtained from a time domain analysis.

The remainder of this paper is laid out as follows: the analyzed time series records of NO_2 and O_3 are summarized in "Air quality data and monitoring stations". "Statistical methods" briefly introduces basic concepts related to the wavelets-based methods. Furthermore, the time series clustering procedure is also described. The results as groups of the monitoring stations extracted on the basis of their corresponding variance/covariance profiles are presented and discussed in "Results and discussion". Finally, the last section is devoted to conclusions.

Air quality data and monitoring stations

A total of 36 monitoring stations measuring NO_2 and O_3 pollutants within Portugal mainland (see Fig. 1) were selected taking into account the efficiency data collection (> 85 %) during the 8-years period

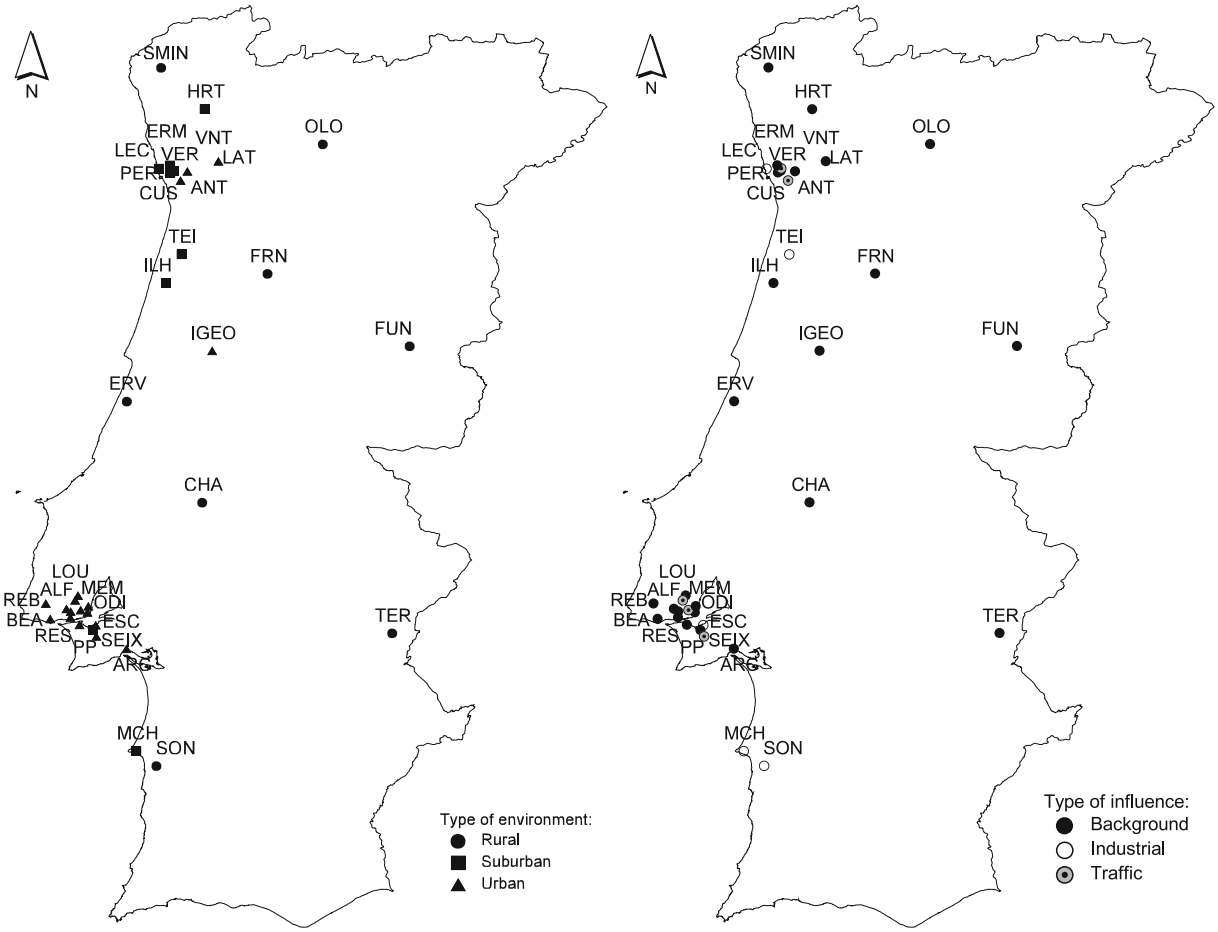

Fig. 1 Spatial location and classification (environment and influence) of the air quality monitoring stations

2005–2013. The data are reported on an hourly basis and is available at the Portuguese Air Quality Database (http://qualar.apambiente.pt/). The quality control of the data is based on common methods and criteria for diagnosing and assessing the ambient air quality in Member States (Directive 2008/50/EC). Before 2005 year, the group of monitoring stations was considerably smaller being that the reason for not considering the previous periods in this study.

The spatial coverage is suitable over Portugal area with at least one station located in each administrative region. Nevertheless, more than 50 % of the monitoring sites are located in the metropolitan areas of Lisbon (36 %) and Oporto (22 %). Because of that, the majority of stations (20 out of 36) are classified as urban environment (see Table 1) with only eight suburban and eight rural. This group of 36 stations also includes the three types of influence: 26 background, five with traffic, and five with industrial influence.

The air quality data used in this work consists of 36 pairs of NO_2 and O_3 time series hourly collected during the 8-years period 2005–2013. As previously referred, the stations were chosen according to the efficiency data collection (> 85 %), and the missing data were replaced by the value from the nearest-neighbor station evaluated from Euclidean distance (Speed 2003).

Figure 2 displays the hourly NO_2 and O_3 levels for two illustrative background stations: SMIN (rural) and ARC (urban). Note that both NO_2 average and variability levels at SMIN station are lower when compared to those collected at ARC station. In general, the high NO_2 values and variability at urban stations are explained by the multiple and different emission sources of NOx existent in urban areas, namely road transport and residential combustion. Regarding the O_3, mean level is slightly higher whereas the variability of O_3 values is lower at SMIN station, which is expected since O_3 is a secondary pollutant formed in the atmosphere along the transport from polluted areas and reaching higher values in remote areas. A closer look to Fig. 2 also indicates that these series exhibit an annual periodicity with the annual minimum NO_2 values being reached simultaneously with the annual maximum O_3 values. The annual periodicity of the NO_2 and O_3 series and the association between them are more notorious at ARC station, and thus, it is expected that the variance/covariance contribution of the annual frequency will be higher in urban stations.

The baseline O_3 fluctuations (and indirectly for the NO_2 precursor) are mainly caused by the seasonal variation of the solar flux (Austin et al. 2007) as well as by other sources with direct impact in the slow trend of the series, such as alterations in deposition due to land use changes (e.g., Emberson et al. 2013). Also, there are other time scales associated with important physical processes. Namely, the O_3 photochemical production cycle is responsible for the typical diurnal profile (intra-day, < 12 h). Furthermore, the daily cycle (24 h) in ground-level O_3 is associated with the day/night variation of the solar flux and the resulting differences between daytime photochemical production and nighttime removal of ozone (Hogrefe et al. 2000). Finally, the synoptic patterns having a prevalence time of about 4–7 days contain fluctuations related to changing weather patterns (Sebald et al. 2000). Moreover, negative cross-correlation between O_3 and its precursor NO_2 is expected due to the chemical atmospheric processes involved (Seinfeld and Pandis 2006).

Statistical methods

This section provides an outline of discrete wavelet analysis (and Maximal Overlap Discrete Wavelet Transform, or MODWT in short) useful in the present setting, referring the reader to the book of Percival and Walden (2006) and the references therein for a more detailed description. Furthermore, this section also describes the Wavelet decomposition-based Clustering (WdC hereafter) approach used in this study to classify bivariate time series of NO_2 and O_3. The purpose of this analysis is twofold: first, wavelet decomposition is applied to NO_2 and to O_3 time series in order to identify the most relevant scales in what concerns to variability and also joint variability for each station. Second, based on such features, the WdC method is applied for grouping stations with similar profiles.

Wavelet-based decomposition and feature extraction

The MODWT is a linear filtering operation which is the base of an additive decomposition of a given time

Table 1 Identification of the Portuguese air quality monitoring stations: coordinates (LAT and LONG) and classification (type of environment and influence)

Station name	Abbrev.	Environment	Influence	LAT	LONG
Alfragide	ALF	Urban	Background	−9.21	38.74
Antas	ANT	Urban	Traffic	−8.59	41.16
Arcos	ARC	Urban	Background	−8.89	38.53
Beato	BEA	Urban	Background	−9.11	38.73
Chamusca	CHA	Rural	Background	−8.47	39.35
Custoias	CUS	Suburban	Background	−8.65	41.21
Entrecampos	ENT	Urban	Traffic	−9.15	38.75
Ermesinde	ERM	Urban	Background	−8.55	41.22
Ervedeira	ERV	Rural	Background	−8.89	39.92
Escavadeira	ESC	Urban	Industrial	−9.07	38.66
Fornelo do Monte	FRN	Rural	Background	−8.10	40.64
Fundao	FUN	Rural	Background	−7.30	40.23
Horto	HRT	Suburban	Background	−8.45	41.57
Int Geofisico Coimbra	IGEO	Urban	Background	−8.41	40.22
Ilhavo	ILH	Suburban	Background	−8.67	40.59
Laranjeiro	LAR	Urban	Background	−9.16	38.66
Centro Laticinios	LAT	Urban	Background	−8.38	41.27
Leca	LEC	Suburban	Background	−8.63	41.22
Loures	LOU	Urban	Background	−9.17	38.83
Monte Chaos	MCH	Suburban	Industrial	−8.89	37.95
Mem−Martins	MEM	Urban	Background	−9.35	38.78
Odivelas	ODI	Urban	Traffic	−9.18	38.80
Olivais	OLI	Urban	Background	−9.11	38.77
Lamas Olo	OLO	Rural	Background	−7.79	41.37
Perafita	PER	Suburban	Industrial	−8.71	41.23
Paio Pires	PP	Suburban	Background	−9.08	38.63
Quinta Marques	QMARQ	Urban	Background	−9.32	38.70
Reboleira	REB	Urban	Background	−9.23	38.75
Restelo	RES	Urban	Background	−9.21	38.71
Alto Seixalinho	SEIX	Urban	Traffic	−9.06	38.65
Senhora Minho	SMIN	Rural	Background	−8.70	41.80
Sonega	SON	Rural	Industrial	−8.72	37.87
Teixugueira/Estarreja	TEI	Urban	Industrial	−8.58	40.75
Terena	TER	Rural	Background	−7.40	38.62
Vermoim	VER	Urban	Traffic	−8.63	41.23
Vila Nova Telha	VNT	Suburban	Background	−8.65	41.25

Stations are presented in alphabetical order of their abbreviations

series X_t, for $t = 1, \ldots, T$ and consists of reexpressing X_t as the sum of $J + 1$ sub-series corresponding to each time-scale, that is

$$X_t = \sum_{j=1}^{J} D_j + S_J,$$

where D_j, for $j = 1, \ldots, J$, represents the time series with wavelet coefficients (details) corresponding to the pass-band filtering scales $\tau_j = 2^{j-1}$ and S_J is a time series with scaling coefficients (smooth) which corresponds to the remaining parcel of the decomposition. The wavelet coefficients D_j at scale $\tau_j =$

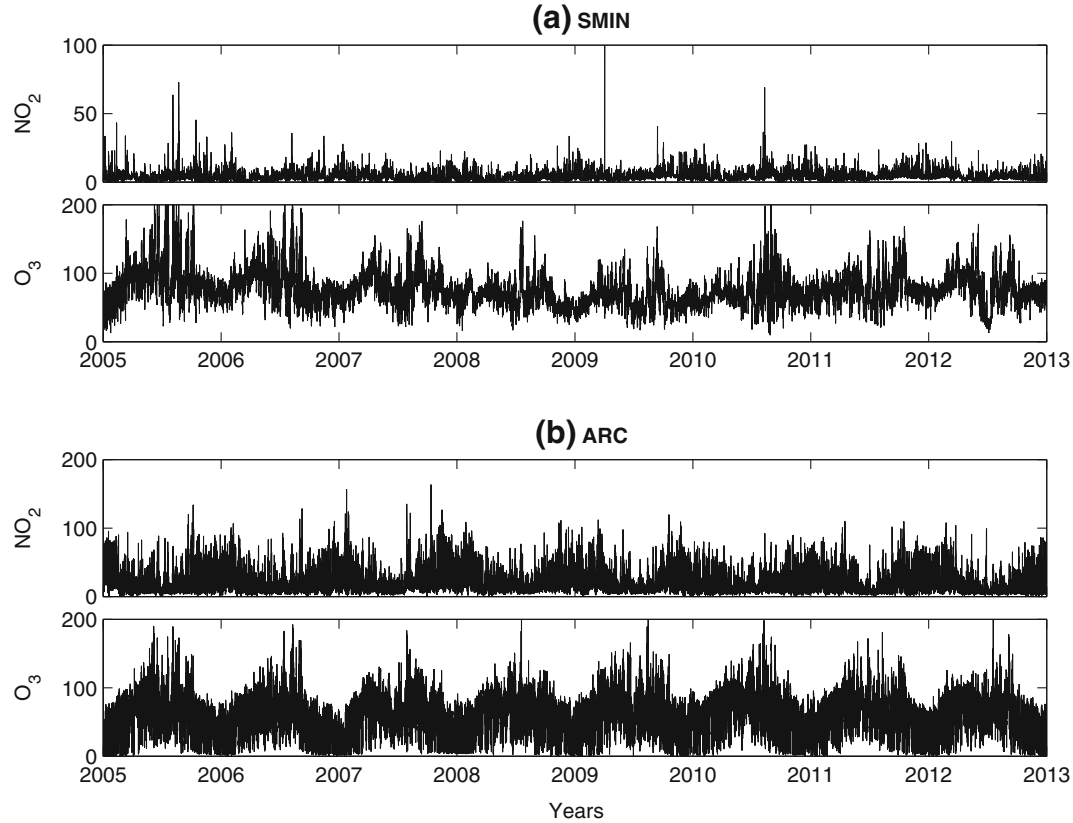

Fig. 2 Hourly concentration values ($\mu g\, m^{-3}$) of NO_2 and O_3 for the background stations **a** rural SMIN and **b** urban ARC

2^{j-1} are associated with frequencies in the interval $\left[1/2^{j+1}, 1/2^{j}\right]$ and thus scale τ_j captures the dynamics over intervals with duration from 2^j to 2^{j+1} time units. Consequently, S_{J} includes information from all scales above 2^J time units.

Let $\{\delta_{j,l} : l = 0, 1, \ldots, L_j\}$ be the MODWT wavelet filter of length L_{j} associated with the scale τ_{j}, where $L_j = (2^j - 1)(L-1) + 1$ and L is the width of the base filter (i.e., for $j = 1$). Further assume that

$$W_{j,t}^{X} = \sum_{l=0}^{L_j - 1} \delta_{j,l} X_{t-l},$$

represents the stochastic process by filtering the discrete time stochastic process X_{t} with the MODWT filter $\delta_{j,l}$. Then, the time independent MODWT wavelet variance at scale τ_j is defined as

$$\nu_X^2(\tau_j) := V\left(W_{j,t}^{X}\right),$$

provided that it exists and is finite. Thus,

$$V(X_t) = \sum_{j=1}^{\infty} \nu_X^2(\tau_j),$$

which implies that the wavelet analysis decomposes the variance of (X_{t}) across wavelet scales. A similar decomposition can be obtained for the covariance between two stochastic processes X_{t} and Y_{t} with MODWT coefficients $W_{j,t}^{X}$ and $W_{j,t}^{Y}$, respectively, defined as

$$Cov(X_t, Y_{\mathrm{t}}) = \sum_{j=1}^{\infty} \nu_{XY}(\tau_{\mathrm{j}}),$$

where $\nu_{\mathrm{XY}}(\tau_{\mathrm{j}})$ is the covariance between $W_{j,t}^{X}$ and $W_{j,t}^{Y}$ wavelet scales. In a bivariate framework, the scale-by-scale wavelet variance/covariance quantification can be rearranged in the following symmetric matrix

$$\mathcal{C}(\tau_j) := \begin{pmatrix} \nu_{\mathrm{X}}^2(\tau_{\mathrm{j}}) & \nu_{XY}(\tau_{\mathrm{j}}) \\ \nu_{\mathrm{YX}}(\tau_{\mathrm{j}}) & \nu_{\mathrm{Y}}^2(\tau_{\mathrm{j}}) \end{pmatrix}.$$

After setting the base filter, $W_{j,t}^{X}$ can be straightforward estimated by considering circular boundary conditions due to the periodic characteristics of NO_2 and O_3 time series. Hence, $\mathcal{C}(\tau_j)$ can be estimated through the unbiased empirical counterpart of its components, namely

$$\hat{\nu}_X^2(\tau_j) := \frac{1}{M_j} \sum_{t=L_j-1}^{T-1} \left(\hat{W}_{j,t}^{X}\right)^2 \tag{1}$$

and

$$\hat{\nu}_{XY}(\tau_j) \equiv \hat{\nu}_{YX}(\tau_j) := \frac{1}{M_j} \sum_{t=L_j-1}^{T-1} \hat{W}_{j,t}^{X} \hat{W}_{j,t}^{Y}, \tag{2}$$

where $M_j = T - L_j + 1$ represents the number of wavelet coefficients excluding the boundary coefficients that are affected by the circular assumption of the wavelet filter.

The wavelet filter is selected to obtain an adequate variance decomposition and variance estimation across scales. In particular, the least asymmetric filter of width $L = 8$, i.e., LA(8), was adopted in this analysis since it yields coefficients that are approximately uncorrelated between scales while having a filter width short enough to minimize the number of boundary coefficients. Furthermore, LA filters exhibit approximately linear phase and thus allowed to align the sinusoidal components in all scales with the original time series by time shift, for visualization purposes. Finally, the number of scales J is restricted to the length of the time series (T) and the filter width (L) as follows

$$J < \log_2\left(\frac{T}{L-1} + 1\right),$$

which resulted in $J \leq 13$ for this study. It is important to remark that MODWT analysis up to $J = 13$ scales includes the frequencies of maximum interest in NO_2 and O_3 time series, i.e., up to annual periodicity.

In this work, the NO_2 and O_3 series were analysed after normalization towards zero mean and unit variance, i.e., $(X_t - m)/s$, where m represents the constant mean and s the standard deviation of the original time series (X_t). Since the MODWT allows partitioning the total variance of the original series by scale, the variance of the normalized series associated with each scale corresponds to the percentage of X_t variance associated with such scale. The normalized series are also used to compute the wavelet covariance as measure of the association between NO_2 and O_3 series across scales.

Clustering of bivariate time series

The clustering procedure builds a hierarchy from the individual elements by progressively merging more similar clusters, using an appropriate dissimilarity measure and a group linkage criterion (see, e.g., Everitt et al. 2011). The dissimilarity matrix, d_w, has entries $d_w(i, i')$ corresponding to the pairwise comparison between objects i and i'. In this work, each object i corresponds to a set of bivariate time series, say $[X_{i,t}\ Y_{i,t}]$. The comparison between each pair of bivariate series i and i' is based on their corresponding wavelet variance/covariance matrices, using the following distance measure proposed by D'Urso et al. (2014),

$$d_w(i, i') = \left\{\left(0.5\, d_{wv}(i, i')\right)^2 + \left(0.5\, d_{wc}(i, i')\right)^2\right\}^{\frac{1}{2}},$$

which weights evenly the components $d_{wv}(i, i')$ and $d_{wc}(i, i')$, connected, respectively, to the wavelet variance and covariance of the time series. Note that $d_{wv}(i, i')$ takes into account the differences in variance across scales for the objects i and i' as

$$d_{wv}(i, i') = \sum_{j=1}^{J} \|diag(\mathcal{C}_i(\tau_j)) - diag(\mathcal{C}_{i'}(\tau_j))\|,$$

where $diag(\mathcal{A})$ denotes the principal diagonal of a matrix $\mathcal{A}$ and $\|\cdot\|$ represents the Euclidean norm. On the other hand, the component

$$d_{wc}(i, i') = \sum_{j=1}^{J} \|\nu_{X_{i,t}Y_{i,t}}(\tau_j) - \nu_{X_{i',t}Y_{i',t}}(\tau_j)\|,$$

quantifies the differences in wavelet covariances across scales. Distances $d_{wv}(i, i')$ and $d_{wc}(i, i')$ are estimated by replacing its components by their empirical counterparts from Eqs. 1 and 2.

Finally, the clustering procedure involves obtaining a dendrogram based on the application of classical cluster techniques to the d_w matrix. In particular, unweighted average distance (average linkage), shortest distance (single), and furthest distance (complete) were considered for the group linkage criterion. The group linkage is chosen as to maximize the dendrogram's goodness-of-fit, evaluated through the cophenetic correlation coefficient between distances matrix

d_w and distances represented in the cophenetic matrix (Everitt et al. 2011, p. 91). The closer the coefficient is to one, the more accurately the clustering procedure reflects the original data.

Results and discussion

In this section, we first present the results of the wavelet-based method for the rural SMIN and the urban ARC monitoring stations. Afterwards, the results are summarized for the remaining stations by their scale-by-scale variance/covariance profiles. The result of the profile-based clustering procedure are represented in a dendrogram.

As previously mentioned, the times series of NO_2 and O_3 series are of non-stationary nature which supports the demand to use MODWT for the variance decomposition per frequency scale. These time series exhibit an annual periodicity with negative NO_2 and O_3 association, being this pattern more visible for the urban station ARC which is related to a more relevant chemical interaction processes between the two pollutants in urban/polluted areas. This annual periodicity can be properly isolated at higher scales in the MODWT analysis, as illustrated in Fig. 3 for the O_3 time series of the ARC station and taking $j = 13$ (i.e., a time scale from 2^{13} to 2^{14} hours which includes the yearly period). Note that, as j decreases, the time scale becomes shorter and therefore the wavelet scales include the higher frequencies of the original signal. Thus, the variance associated to each scale represents the parcel of the original variance within a certain frequency interval. In addition to τ_{13}, also wavelet scales τ_3 and τ_4 exhibit high O_3 variances. Such scales correspond to time periods between 8 to 16 h (intra-diurnal) and 16 to 32 h (daily), respectively. The same scales have also a relevant contribution for the NO_2 variance (see Fig. 3b). Note that the negative association between NO_2 and O_3 values is particularly obvious for the wavelet scale including the annual periodicity ($j = 13$ in Fig. 3a, b). The importance of the daily time scales of the ARC urban station is justified by the more relevant chemical transformations occurring in areas with high NO_2 emissions (urban areas), which are responsible for high variability on daily profiles: the intra-diurnal (τ_3) related with the traffic profile and the higher night/day magnitude differences (τ_4) due to the photochemical processes involved. On the other hand, rural stations are mainly influenced by transport processes and not directly affected by primary pollutant emissions, which justify the less variability found.

Spectral analysis is commonly used to describe a time series in the frequency domain and to quantify the contribution of each frequency to the variance of the series. In particular, the highest spectrum amplitudes localize the most important cyclic components in each series and the area of each component corresponds to the contribution of the variance associated with that frequency band to the total variance of the original time series. In this sense, wavelet scales may be interpreted as the equivalent to the spectral density function (the variance as a function of the frequency content of the series) but regarding frequency intervals instead of single frequencies.

To help the comparison between SMIN and ARC profiles, Fig. 4a–d shows the spectrum of the NO_2 and O_3 time series as well as the spectrum of the wavelet scales exhibiting high variance (i.e., $W_{j,t}$ for $j = 3, 4, 6, 13$). Note that the components identified in NO_2 are placed in similar frequency locations for both stations, clearly indicating the presence of the same periodicities in both stations. Regarding the weight of each frequency component, the contribution of τ_6 is similar in both stations. Furthermore, τ_3 and τ_{13} contributions are higher for ARC station than for SMIN station. Similar remarks can be depicted from the spectral analysis in O_3 regarding τ_3 and τ_{13} contributions. Urban ARC site exhibits a much larger τ_4 contribution in comparison with rural SMIN station.

Figure 4e–f displays the NO_2 and O_3 cross-spectrum for SMIN and for ARC stations (respectively), illustrating the decomposition of NO_2 and O_3 covariance as a function of frequency. Note that covariance is larger for the urban ARC station when considering all wavelet components which is justified by the above mentioned chemical processes transformation (interconversion) involving both species.

The clustering of the stations considered the corresponding variance/covariance profiles. The results show that the group linkage criterion with the highest cophenetic correlation coefficient is the average linkage, i.e., the clustering approach merging stations and clusters of stations whose average distance with respect to all pairs is minimized. The coefficient value equals 0.86 indicating that the clustering is quite fit to the original data. Figure 5 represents

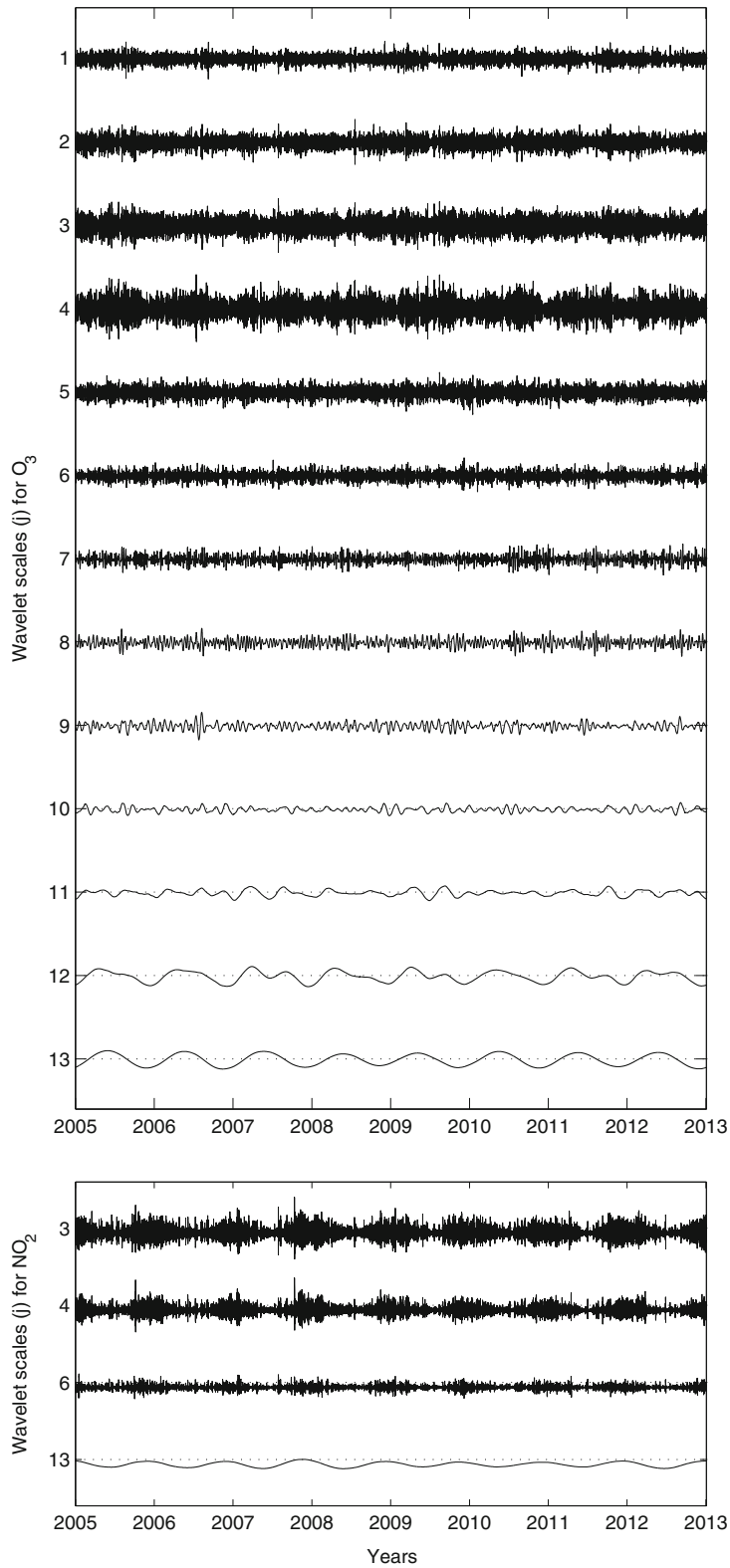

Fig. 3 Wavelet scales of the hourly O_3 (top) and NO_2 (*bottom*) values for the urban background station ARC

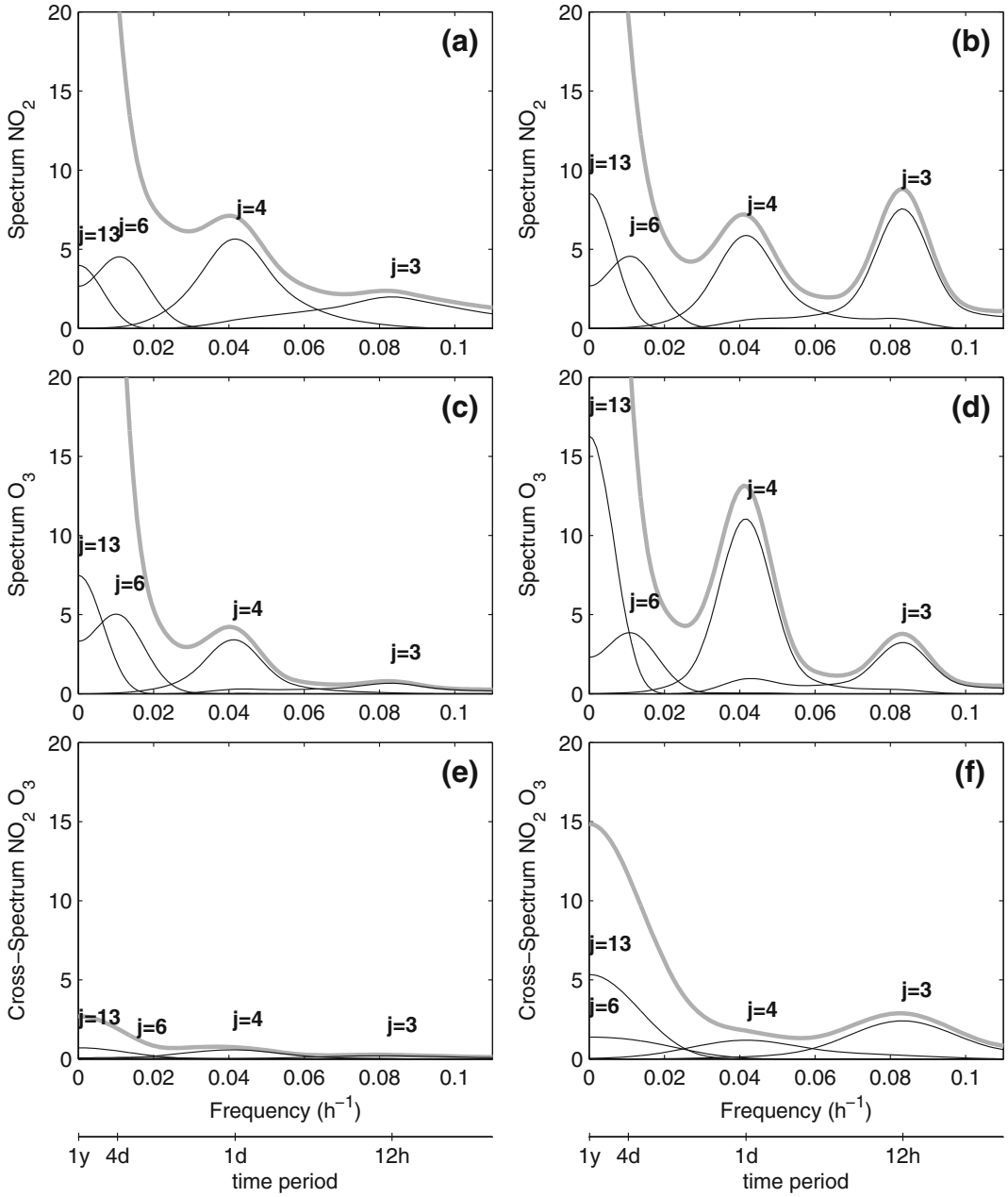

Fig. 4 Spectrum and cross-spectrum of hourly NO_2 and O_3 values for the background stations **a, c, e** rural SMIN and **b, d, f** urban ARC. Spectrum and cross-spectrum of the original time series (*grey*) and of $W_{j,t}$ scales for $j = 3, 4, 6, 13$ (*black*) estimated via Blackman-Tukey algorithm

the dendrogram showing how stations are linked: the lower the linkage level (distance), the higher their similarity. The dendrogram clearly distinguishes three groups of similar monitoring stations with respect to variance/covariance profile (i.e., lowest distance).

Class A (blue) typically contains urban stations with the exception of PP (background). Class B (green) includes suburban (PER, VNT, CUS, LEC, HRT, and ILH), urban (VER, ANT, IGEO, TEI, and LAT), and also some rural stations. Finally, class C

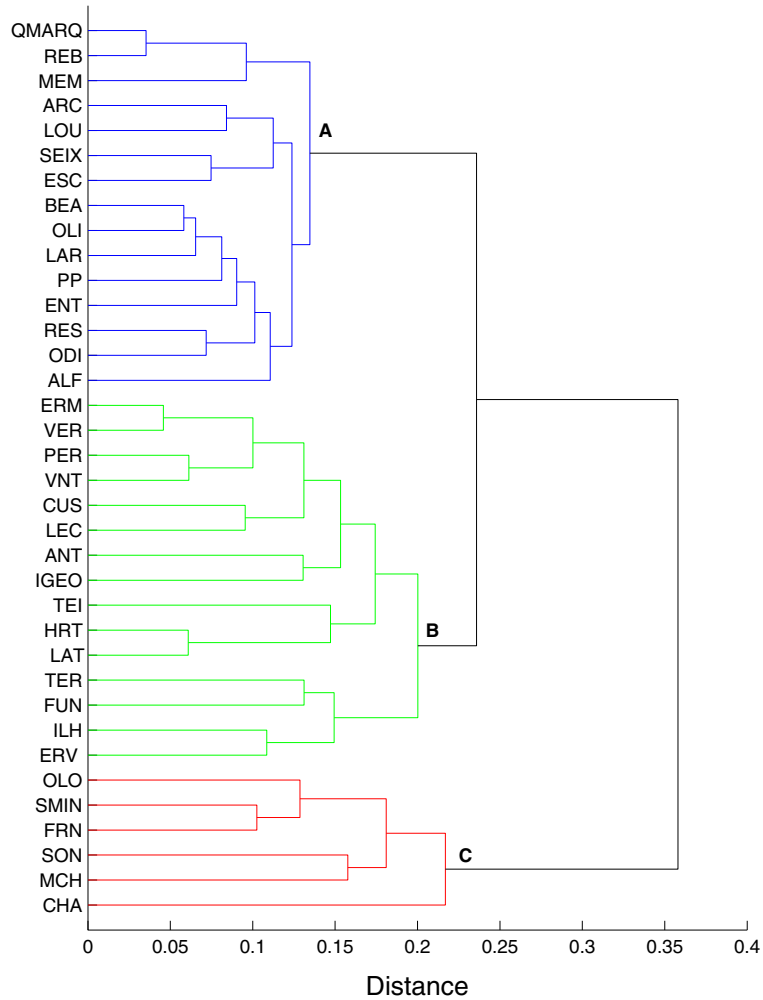

Fig. 5 Dendrogram showing the hierarchical clustering of the monitoring stations (average linkage criterion)

(red) includes rural stations (except suburban MCH). It is worthwhile to mention that the stations in class C are highly heterogeneous (reflecting existing differences in variance/covariance patterns among rural stations and higher geographical distance) when compared with those in classes A and B. A more detailed analysis shows that a geographical pattern exist, namely in urban stations: the stations located in Lisbon area belong to class A whereas all urban stations located in Porto/North region (more influenced by traffic and industrial activity) fall into class B.

Aiming for a more comprehensive interpretation of the three groups identified in the dendrogram, Fig. 6 shows the NO_2 and the O_3 variance/covariance contribution per wavelet scale and per station, where darker colors indicate higher contribution.

As observed in Fig. 6a, monitoring stations in class A and class B exhibit the highest contribution for the NO_2 total variance in τ_3 and τ_4 scales whereas for monitoring stations in class C the total variance is more spread over all scales. With respect to O_3 variance in Fig. 6b, the scale τ_4 is of major importance in class A and in class B stations, being higher for stations in the latter class. This result is related to the higher amplitude and variability of the diurnal cycle of both pollutants, which is directly influenced by NO_2 emissions and O_3 chemistry (production and consumption). Finally, Fig. 6c shows that NO_2 and O_3 are negatively associated in the three classes and for all scales. Note that the time series of NO_2 and O_3 are more associated in scales τ_3, τ_4, τ_{13} for stations in class A, and τ_3, τ_4 in class B. The stations in class C

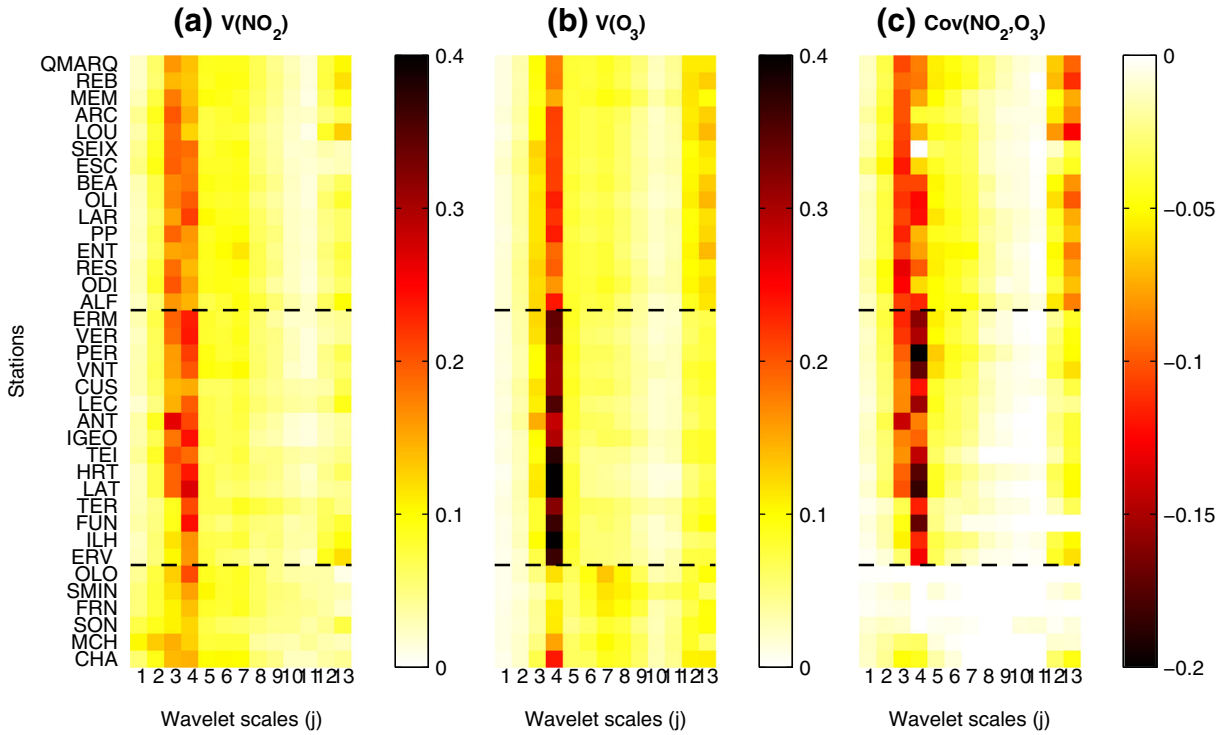

Fig. 6 Mosaic plot representing the contribution to the total variance/covariance of the time series per wavelet scale for all stations (*darker colors* indicate higher absolute values). *Dashed lines* delimitate the three clusters represented in Fig. 5 with stations following the same order

are those with lowest association between NO_2 and O_3 across scales. Chemical processes involving these two species are more relevant in polluted environments (urban and suburban areas) which explain the higher absolute association and correlation between them. Since this chemistry involves the reaction of NO_2 and the production of O_3, the negative cross correlation is justified.

The subgroup of TER, FUN, ILH, and ERV stations in class B exhibits a similar profile to that of the remaining stations within this class, although also sharing some characteristics with the rural stations in class C. In particular, for NO_2 variance (Fig. 6a) and for O_3 and NO_2 covariance (Fig. 6c), the contribution of τ_4 resembles to that of the remaining stations in class B whereas the contribution of τ_3 is more similar to that of stations in class C. This suggests that these particular stations are located near rural environments but with more high amplitude and variability of the (concentration) daily profiles.

In summary, these results allow to better understand the characteristics associated to each monitoring station and to identify similarities/differences among them. In what concern to its classification in terms of environment and influence, some exceptions were identified; namely, the PP suburban station which profile resembles more the urban stations; TEI and LAT urban stations that show a suburban behavior; and also the rural stations TER, FUN, and ERV with suburban characteristics. Finally, it is also important to mention that two different classified stations—rural SON and suburban MCH (both industrial)—although being geographically close, exhibit different profiles as observed by the high linkage level at which these stations merge into one group.

Conclusions

In this paper, we investigate the variance/covariance profile of 36 monitoring sites, measuring ozone (O_3) and nitrogen dioxide (NO_2) hourly concentrations collected in Portugal mainland. The resulting individual profiles are embedded in a wavelet decomposition-based clustering algorithm to identify groups of stations exhibiting similar profiles. The results of the

cluster analysis identify three groups of monitoring stations; one mainly containing urban stations, another including suburban, urban, and some rural stations, and a third class mainly formed by rural stations. The results also indicate a geographical pattern among the urban stations. For both pollutants, intra-diurnal and daily time scales exhibit the highest variance in particular for the urban stations, which is justified by the more relevant chemical activity occurring in areas with high NO_2 emissions (urban areas), responsible for high variability on daily profiles. Such chemical processes are also the reason why NO_2 and O_3 are highly negatively cross-correlated in suburban and urban sites as compared with the rural ones. This study also identified some sites which need further revision with respect to their classification according to the type of environment and influence. This group of stations includes the PP suburban station, TEI and LAT urban industrial stations and the rural stations TER, FUN and ERV with suburban characteristics.

Acknowledgments This work was supported by Portuguese Funds through FCT - Foundation for Science and Technology, in the context of the projects UID/CEC/00127/2013 and Incentivo/EEI/UI0127/2014 (IEETA/UA, Instituto de Engenharia Electrónica e Informática de Aveiro, www.ieeta.pt) and UID/MAT/04106/2013 (CIDMA/UA, Centro de I&D em Matemática e Aplicações, www.cidma.mat.ua.pt). S. Gouveia acknowledges the postdoctoral grant by FCT (ref. SFRH/BPD/87037/2012), financed through POPH - QREN programme (European Social Fund and Nacional funds). Andres M. Alonso acknowledges the support of CICYT (Spain) Grants ECO2011-25706 and ECO2012-38442. The authors also gratefully acknowledge to the Portuguese Environmental Agency for providing the air quality monitoring data.

Conflict of interests The authors declare that they have no conflict of interest.

References

Adame, J.A., Bolívar, J.P., & De la Morena, B.A. (2010). Surface ozone measurements in the southwest of the Iberian Peninsula (Huelva, Spain). *Environmental Science and Pollution Research International*, *17*, 355–368.

Alkuwari, F.A., Guillas, S., & Wang, Y. (2013). Statistical downscaling of an air quality model using Fitted Empirical Orthogonal Functions. *Atmospheric Environment*, *81*, 1–10.

Alonso, A.M., Berrendero, J.R., Hernández, A., & Justel, A. (2006). Time series clustering based on forecast densities. *Computational Statistics and Data Analysis*, *51*, 762–776.

Austin, J., Hood, L.L., & Soukharev, B.E. (2007). Solar cycle variations of stratospheric ozone and temperature in simulations of a coupled chemistry-climate model. *Atmospheric Chemistry and Physics*, *7*, 1693–1706.

Carvalho, A., Monteiro, A., Ribeiro, I., Tchepel, O., Miranda, A.I., Borrego, C., Saavedra, S., Souto, J.A., & Casares, J.J. (2010). High ozone levels in the northeast of Portugal: analysis and characterization. *Atmospheric Environment*, *44*, 1020–1031.

Clapp, L.J., & Jenkin, M.E. (2001). Analysis of the relationship between ambient levels of O_3, NO_2 and NO as a function of NOx in the UK. *Atmospheric Environment*, *35*, 6391–6405.

De Iaco, S. (2011). A new space-time multivariate approach for environmental data analysis. *Journal of Applied Statistics*, *38*, 2471–2483.

D'Urso, P., & Maharaj, E.A. (2012). Wavelets-based clustering of multivariate time series. *Fuzzy Sets and Systems*, *193*, 33–61.

D'Urso, P., De Giovanni, L., & Maharaj, E.A. (2014). Wavelet-based self-organizing maps for classifying multivariate time series. *Journal of Chemometrics*, *28*, 28–51.

Everitt, B.S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*. Chichester: Wiley.

Emberson, L.D., Kitwiroon, N., Beevers, S., Buker, P., & Cinderby, S. (2013). Scorched Earth: How will changes in the strength of the vegetation sink to ozone deposition affect human health and ecosystems? *Atmospheric Chemistry and Physics*, *13*, 6741–6755.

Figueiredo, M.L., Monteiro, A., Lopes, M., Ferreira, J., & Borrego, C. (2013). Air quality assessment of Estarreja, an urban industrialized area, in a coastal region of Portugal. *Environmental Monitoring and Assessment*, *185*, 5847–5860.

Finazzi, F., Scott, E.M., & Fassó, A. (2013). A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. *Applied Statistics*, *62*, 287–308.

Fiore, A.M., Jacob, D.J., Mathur, R., & Martin, R.V. (2003). Application of empirical orthogonal functions to evaluate ozone simulations with regional and global models. *Journal of Geophysical Research*, *108*, D14,443.

Hogrefe, C., Rao, S.T., Zurbenko, I.G., & Porter, P.S. (2000). Interpreting the information in ozone observations and model predictions relevant to regulatory policies in the eastern United States. *Bulletin of the American Meteorological Society*, *81*, 2083–2106.

Ignaccolo, R., Ghigo, S., & Giovenali, E. (2008). Analysis of air quality monitoring networks by functional clustering. *Environmetrics*, *19*, 672–686.

Im, U., Incecik, S., Guler, M., Tek, A., Topcu, S., Unal, Y.S., Yenigun, O., Kindap, T., Talat Odman, M., & Tayanc, M. (2013). Analysis of surface ozone and nitrogen oxides at urban, semi-rural and rural sites in Istanbul, Turkey. *Science of the Total Environment*, *443*, 920–931.

Joly, M., & Peuch, V.H. (2012). Objective classification of air quality monitoring sites over Europe. *Atmospheric Environment*, *47*, 111–123.

Kracht, O., Reuter, H.I., & Gerboles, M. (2013). A tool for the SpatioTemporal screening of AirBase Datasets for

abnormal values, European Commission Report 25787 EN, Joint Research Centre.

Kracht, O., Reuter, H.I., & Gerboles, M. (2014). First evaluation of a novel screening tool for outlier detection in large scale ambient air quality datasets. *International Journal of Environment and Pollution*, *55*, 120–128.

Levy, I., Mihele, C., Lu, G., Narayan, J., & Brook, J.R. (2014). Evaluating multipollutant exposure and urban air quality: pollutant interrelationships, neighborhood variability, and nitrogen dioxide as a proxy pollutant. *Environmental Health Perspectives*, *122*, 65–72.

Li, L., Wu, J., Ghosh, J.K., & Ritz, B. (2013). Estimating spatiotemporal variability of ambient air pollutant concentrations with a hierarchical model. *Atmospheric Environment*, *71*, 54–63.

Liu, S., Maharaj, E.A., & Inder, B. (2014). Polarization of forecast densities: a new approach to time series classification. *Computational Statistics and Data Analysis*, *70*, 345–361.

O'Leary, B.F., & Lemke, L.D. (2014). Modeling spatiotemporal variability of intra-urban air pollutants in Detroit: a pragmatic approach. *Atmospheric Environment*, *94*, 417–427.

Monjardino, J., Ferreira, F., Mesquita, S., Perez, A.T., & Jardim, D. (2009). Air quality monitoring: establishing criteria for station classification. *International Journal of Environment and Pollution*, *39*, 321–32.

Monteiro, A., Strunk, A., Carvalho, A., Tchepel, O., Miranda, A.I., Borrego, C., Saavedra, S., Rodriguez, A., Souto, J., Casares, J., & Elbern, H. (2012a). Investigating a high ozone episode in a rural mountain site. *Environmental Pollution*, *162*, 176–189.

Monteiro, A., Carvalho, A., Ribeiro, I., Scotto, M.G., Barbosa, S., Alonso, A., Baldasano, J.M., Pay, M.T., Miranda, A.I., & Borrego, C. (2012b). Trends in ozone concentrations in the Iberian Peninsula by quantile regression and clustering. *Atmospheric Environment*, *56*, 184–193.

Percival, D.B., & Walden, A.T. (2006). *Wavelet methods for time series analysis*. Cambridge: Cambridge University Press.

Reich, B., Cooley, D., Foley, K., Napelenok, S., & Shaby, B. (2013). Extreme value analysis for evaluating ozone control strategies. *Annals of Applied Statistics*, *7*, 739–762.

Rojas, A.L.P., & Venegas, L.E. (2013). Spatial distribution of ground-level urban background O3 concentrations in the Metropolitan Area of Buenos Aires, Argentina. *Environmental Pollution*, *183*, 159–165.

Shaddick, G., & Wakefield, J. (2002). Modelling daily multivariate pollutant data at multiple sites. *Applied Statistics*, *51*, 351–372.

Sharma, D., & Kulshrestha, U.C. (2014). Spatial and temporal patterns of air pollutants in rural and urban areas of India. *Environmental Pollution*, *195*, 276–281.

Sebald, L., Treffeisen, R., Reimer, E., & Hies, T. (2000). Spectral analysis of air pollutants. Part 2: ozone time series. *Atmospheric Environment*, *34*, 3503–3509.

Seinfeld, J.H., & Pandis, S.N. (2006). *Atmospheric Chemistry and Physics: from air pollution to climate change*, 2nd Edition. New York: Wiley.

Shi, P., Xie, P.-H., Qin, M., Si, F.-Q., Dou, K., & Du, K. (2014). Cluster analysis for daily patterns of SO_2 and NO_2 measured by the DOAS system in Xiamen. *Aerosol and Air Quality Research*, *14*, 1455–1465.

Speed, T. (2003). *Statistical Analysis of Gene Expression Microarray Data*. Boca Raton: CRC Press.

Statheropoulos, M., Vassiliadis, N., & Pappa, A. (1998). Principal component and canonical correlation analysis for examining air pollution and meteorological data. *Atmospheric Environment*, *32*, 1087–1095.

Vilar, J.A., Alonso, A.M., & Vilar, J.M. (2010). Nonlinear time series clustering based on non-parametric forecast densities. *Computational Statistics and Data Analysis*, *54*, 2850–2865. http://www.sciencedirect.com/science/article/pii/S016794730900067X.

Tchepel, O., Costa, A.M., Martins, H., Ferreira, J., Monteiro, A., Miranda, A.I., & Borrego, C. (2010). Determination of background concentrations for air quality models using spectral analysis and filtering of monitoring data. *Atmospheric Environment*, *44*, 106–114.